Data Clustering and Analyzing Techniques Using Hierarchical Clustering Method

Wen Hu and Qing he Pan

School of Computer and Information Engineering,
Harbin University of Commerce, 150028
570749130@qq.com

Abstract

Data clustering and analyzing techniques are studied by using hierarchical clustering method. A matrix of words is constructed with a randomly chosen RSS list. By collecting data from this list a matrix is built. In the matrix each row corresponds to a article and each column represents a word. Based on the matrix a hierarchical clustering algorithm is designed. In this algorithm the Pearson correlation coefficient is used to compute the distances among different contents. The dendrogram is used to describe the hierarchical relationship of contents and words. And the 2-D graph also is used to represent the dendrogram in another format.

Keywords: *RSS, hierarchical clustering, Pearson correlation coefficient*

1. Introduction

RSS means "Really Simple Syndication", "RDF (Resource Description Framework) Site Summary" or "Rich Site Summary" [1]. Actually all the three explanations indicate the same syndication technique. Now RSS is being wildly used in online news channels, blog and wiki. Using the RSS export of the website user can subscribe to the news and quickly obtain information.

It is an interesting work to collect and classify information provided by RSS. For example we can classify blogs using RSS they provide to find out the similar writing styles or themes and dig out the similar opinions. Businessmen who run electronic commerce websites can identify who are most possible competitors by analyzing and clustering the RSS of other similar websites and make appropriating decisions.

Data clustering [2, 3] is a main method to finish this work. It a technique belongs to unsupervised learning techniques that are different from supervised learning techniques such as neural networks, decision trees and so on. The aim of data clustering is not to training with samples having right answers but to find out certain structures in object data.

In this paper we study the problem of RSS content clustering by using hierarchical clustering method [4]. In Section 2 we describe the format of the RSS data sets. In Section 3 the hierarchical clustering is described and an algorithm is given. In Section 4 an experiment is implemented using the algorithm. In Section 5 we conclude this paper.

2. The data format

In order to execute the data clustering method the first step is to collect the RSS data and store them in certain way. In this research we collect data and store them in matrix. For a RSS url we first obtain all its content data. Then we strip all non-alphanumeric characters, divide the data into separate word and count the times that each word occurring. For example, there

are six articles from a1 to a6 and the titles are "Neowin", "BetaNews" , "digitallyOBSESSED.com DVD Reviews", "Entertainment News Headlines - Yahoo! News", "Softpedia - Windows - All" and "Sunbelt Software New Products" respectively. For these articles we can list common words in this example six words are chosen. So the matrix format is like Table 1.

Table 1. The format of matrix segment

	how	series	service	and	had	has
a1	1	1	2	18	0	9
a2	4	2	3	59	2	8
a3	0	0	0	0	0	0
a4	0	0	0	14	0	10
a5	0	0	0	6	0	0
a6	1	0	2	26	1	2

It is easy to extract the words using regular expression. In order to deal with data easily all words are converted to lowercase. Based on this basic data format we can execute computation with appropriating method. In this paper we use hierarchical clustering method to analyzing data. With using this method the similar of different RSS is computed and compared.

We execute hierarchical clustering algorithm on data that has the format like Table 1. It is obvious that algorithm analyzes the relationship between rows and it is our goal. But using a matrix transposition (like Table 2) the algorithm can be applied to analyze and cluster the "words" without modification. Sometimes it is interesting to understand which words are used together more often than others.

Table 2. The transposition of Matrix in Table1

	a1	a2	a3	a4	a5	a6
how	1	4	0	0	0	1
series	1	2	0	0	0	0
service	2	3	0	0	0	2
and	18	59	0	14	6	26
had	0	2	0	0	0	1
has	9	8	0	10	0	2

In this paper we focus on the format in Table 1 and cluster different RSS contents.

3. Hierarchical Clustering

The hierarchical clustering method will always combine two most similar groups into a single group and construct a new hierarchical structure. In this research each group is composed of elements and each element is a RSS. The algorithm will compute the distances between every pair of groups and combine the two groups into a new group until there is only one group. This process can be depicted in Figure 1.

Figure 1 depicts the process of hierarchical clustering. Suppose that the process starts from A. In the step1 we computes all distance between A and all other elements and the result indicates the distance between A and B is smallest, so A and B is clustered. With the same

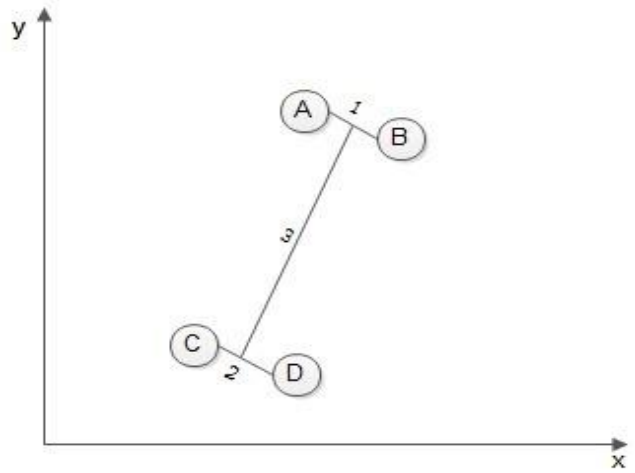

Figure 1. The process of Hierarchical Clustering

in step 2 C and D is clustered. Also in step 3 the groups formed by A B and C D are clustered together.

The distance that we use to evaluate the difference between groups can be computed by Euclidean distance, Pearson correlation coefficient, Manhattan distance, Kendall's (tau) distance or other methods. For example, the Euclidean distance can be computed by formula(1),

$$d = ((x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2)^{\frac{1}{2}} = (\sum_{i=1}^{n} (x_i - y_i)^2)^{\frac{1}{2}} \quad (1)$$

Using this distance we can compare the similarity between two items.

In this research we choose Pearson correlation coefficient to evaluate the difference. The reason is that the different RSS has different number of words and Pearson correlation coefficient can correct this problem since it judges how well two different data sets fit onto a straight line[5]. The formula of Pearson correlation coefficient is given by (1) [6],

$$p(X,Y) = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{(\sum X^2 - \frac{(\sum X)^2}{N})(\sum Y^2 - \frac{(\sum Y)^2}{N})}} \quad (2)$$

In formula (1) X and Y represent two different vectors and in our research each corresponds to one row data in matrix. For example in matrix depicted in Table 1 X may described by vector (1, 4, 0, 0, 0, 1) that is the numbers of words contained in a1.

The hierarchical clustering algorithm can be designed by formula (2). We introduce some symbols in order to describe this algorithm. The first is the data matrix. We can use a vectors list to represent the matrix. The data in table1 are extracted from an experiment. In table1 for each article only six words "how, series, service, and, had, has" are listed but actually there are more words for each article. How to choose words for each article to compare is a problem. In Table 1 the word "and" is used in almost every article and its numbers obviously are bigger. But for word "series" the condition is opposite so there should be a standard to choose words to compare. In this research we set a threshold to choose words. For example the lower limit can be set 20% and upper limit can be set 60%, so for a word if its percent in

all words falls in [0.2, 0.6] then it will be chosen. Let M is the matrix so M may be in the following form according to Table 1,

[(1, 1, 2, 18, 0, 9),
(4, 2, 3, 59, 2, 8),
(0, 0, 0, 0, 0, 0),
(0, 0, 0, 14, 0, 10),
(0, 0, 0, 6, 0, 0),
(1, 0, 2, 26, 1, 2)]

Let n represent the vector number of M.

Algorithm hierarchical clustering(HC)

```
1   Input the data matrix M.
2   While n>1 :
2.1    Look for minimum of (P(m_i, m_j)), i !=j and m_i, m_j ∈ M. and set
P(m_a, m_b) = min(P(m_i, m_j))
2.2    d_min = P(m_a, m_b)
2.3    Compute average value of m_a and m_b, avg(m_a, m_b)
2.4    Use avg (m_a, m_b) to form new cluster,
          new_cluster, m_a is its left child and m_b is its right child.
2.5    Delete m_a and m_b from M
2.6    Append new_cluster to M
2.7    Set n=len(M)
3   return M[0]
```

In step 1 the algorithm accepts M as input. The step 2 is a while loop and the condition is n>1. As definition above n is the length of M or the number of vectors in M. In step 2.1 Pearson correlation coefficient is used compute the distances between every pair of vectors in M and gets the minimum. The minimum is assigned to dmin in step 2.2. In step 2.3 and 2.4 the average value of ma and mb is computed and the new_cluster is formed by using this average value. In 2.5 and 2.6 the ma and mb is deleted from M and the new_cluster is added to M. In step 2.7 n get the new value equaling the length of M. In step 3 the M[0] is returned. We can recursively search M[0] and reconstruct all clusters and children.

4. Experiment

In this section we do real data clustering using HC algorithm. The experiment has three main steps. The first step is to collect RSS data to construct the M matrix. The second step is to apply HC algorithm to M. The third step is to draw the hierarchical structure using the result in the second step.

RSS data is collected by Universal Feed Parser [7], a Python lib. With it we can parse RSS or other feeds including Atom, RDF, RSS, and CDF feed formats. Using it is easy, for example it can be used in following method. First it needs to enter Python environment and next we just input some instruction.

```
>>> import feedparser

>>> d = feedparser.parse('http://www.xxx.com/rss')
```

```
>>> d['feed']['title']

u"XX's Blog"
```

In this research we gather 25 RSS. The Table 3 shows content of these RSS.

Table 3. The content of RSS

RSS	Content
RSS1	Neowin
RSS2	BetaNews
RSS3	digitallyOBSESSED.com DVD Reviews
RSS4	Entertainment News Headlines - Yahoo! News
RSS5	Softpedia - Windows - All
RSS6	Sunbelt Software New Products
RSS7	Softpedia News - Global
RSS8	Health News Headlines - Yahoo! News
RSS9	News
RSS10	Sunbelt Software Updates
RSS11	The Google Weblog
RSS12	SofoTex Software Downloads
RSS13	ExtremeTech
RSS14	About Shareware/Freeware
RSS15	BBC News - Home
RSS16	Help Net Security - Windows Software
RSS17	MajorGeeks.com
RSS18	CNET Download.com 25 Newest Windows Titles
RSS19	OSNews
RSS20	Yahoo! News - Latest News & Headlines
RSS21	IEBlog
RSS22	Odd News Headlines - Yahoo! News
RSS23	Reviews Tom's Hardware
RSS24	Channel 9
RSS25	PCMag.com: New Product Reviews

All these RSS come from different areas and their contents are different. So it is interesting to find out the relations of them.

For each RSS we collect its data, use these data construct M matrix and save M as a .txt file. After extracting all RSS information the text file containing M may look like Figure 2. In Figure 2 we can see some zeros. Zero means there is no this word in corresponding article. But the word whose count equals 0 is also chosen to compute and compare since for words of all articles its percent is in the threshold interval that we set. So though there exists no the word in one article we also include it in M matrix and use it to compute similarity.

Apply HC to M and get the tree-like graph in Figure 3. This tree hierarchy graph is constructed recursively. The contents that are close to each other are joined early and the nodes that have more difference are joined late. In this method a new cluster is created by combining two close clusters. This process continues until there is only one cluster. Usually we call it hierarchical cluster and it will give a dendrogram. The basic steps are:

1) computing all distances between all clusters;

2) Finding out the closest clusters and forming a new cluster;

3) Repeating 1) and 2) until only one cluster exists.

For example we can see "Health News Headlines - Yahoo! News" and "Yahoo! News - Latest News & Headlines" are close. And "digitallyOBSESSED.com DVD Reviews" is far from others.

Blog	how	series	service	and	had	has	people	for
Neowin	1	1	2	18	0	9	0	7
BetaNews	4	2	3	59	2	8	3	28
digitallyOBSESSED.com DVD Reviews	0	0	0	0	0	0	0	0
Entertainment News Headlines - Yahoo! News	0	0	0	14	0	10	3	17
Softpedia - Windows - All	0	0	0	6	0	0	0	8
Sunbelt Software New Products	1	0	2	26	1	2	1	12
Softpedia News - Global	0	3	1	49	1	21	2	27
Health News Headlines - Yahoo! News	1	0	1	46	6	16	3	34
News	6	2	13	92	1	7	2	75
Sunbelt Software Updates	0	0	0	0	0	0	0	0
The Google Weblog	2	0	0	14	0	4	0	9
SofoTex Software Downloads	0	0	0	7	0	0	0	4
ExtremeTech	2	0	0	23	1	1	0	2
About Shareware/Freeware	0	0	0	8	0	2	0	4
BBC News - Home	7	2	0	27	1	4	5	16
Help Net Security - Windows Software	0	1	1	13	0	0	1	4
MajorGeeks.com	0	3	0	27	0	1	0	7

Figure 2. The M Matrix Segment in a Text File

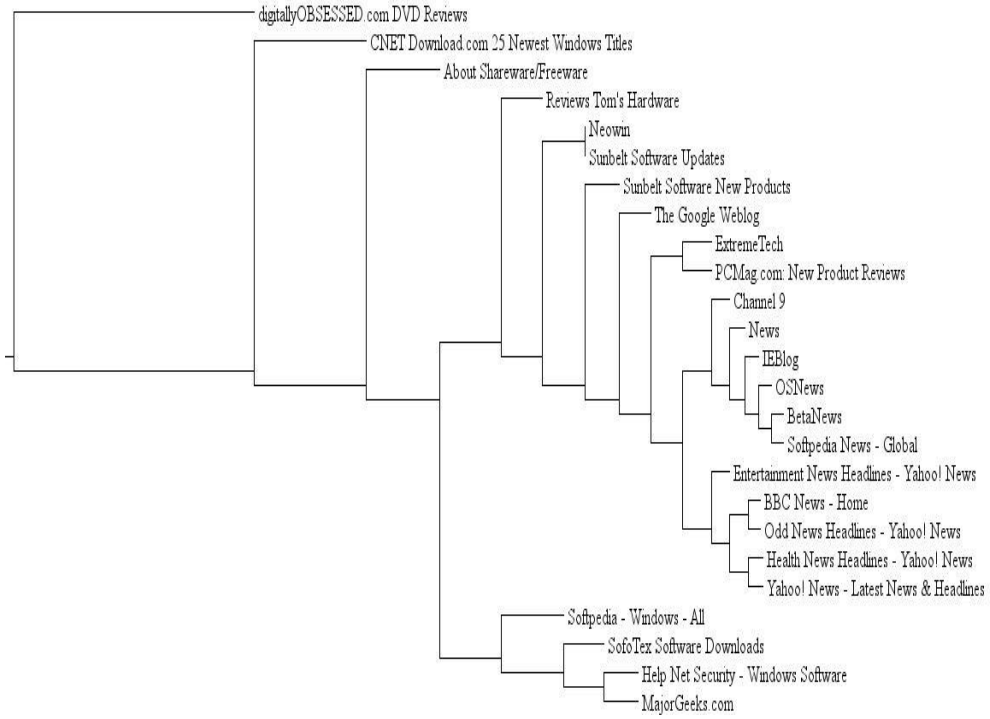

Figure 3. The tree-like graph of 25 RSS feeds clustering

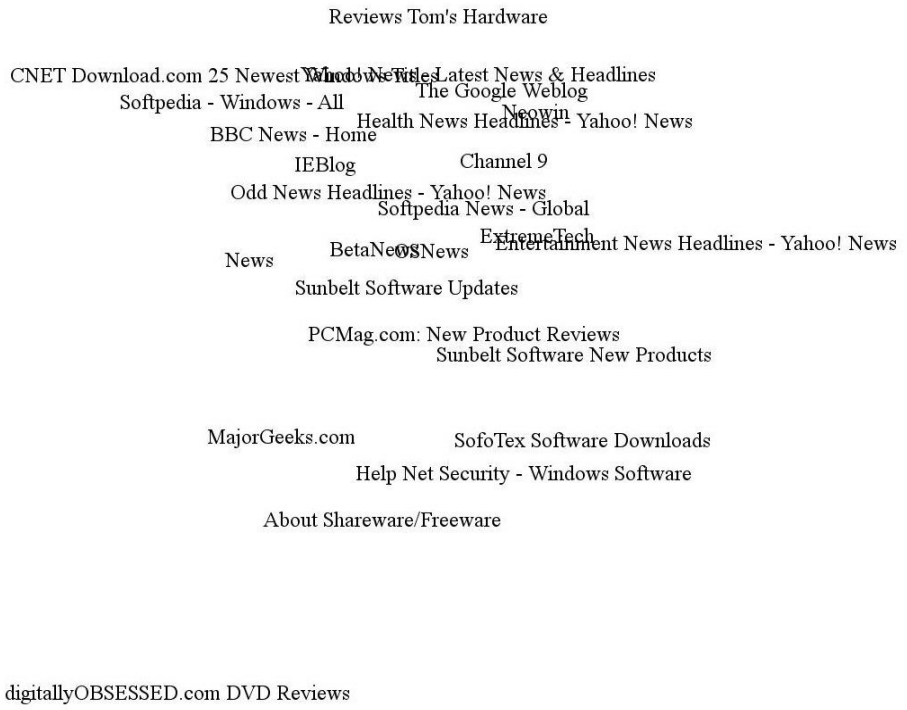

Figure 4. The Two Dimensions Description of 25 RSS

Usually we want to describe the differences between different RSS contents. The difference can be valued by "distance" and the meaning is clear when it is on 2-D plane. In Figure 4 we can view the data in two dimensions format. The content of "digitallyOBSESSED.com DVD Reviews" is obviously different from the other contents. Though some text is overlapping the distances among them make it easy to understand the difference of contents.

In the Section 2 the Table 2 gives the transposition of matrix in table1. In the transposition of matrix the rows represent "words" so when it applies HC to this matrix the results give the relations among words. The dendrogram of words cluster will help us understand which words we will use together among elected articles. In Figure 5 the whole dendrogram of all words of 25 RSS feeds is given. In this figure 129 words are clustered. These words are chosen by rule mentioned above. Because there are more words to cluster than RSS feeds the patterns and trends of words clustering are not so clear when compared with RSS clustering. Also we can see in all 129 words the nouns occupy very small fraction but adjectives, adverbs and pronouns occupy very big fraction. One reason may be that all 25 RSS feeds are not very similar. Another reason may be the thresholds we set to choose words. When the thresholds vary the number of words will change and the different dendrogram will be produced.

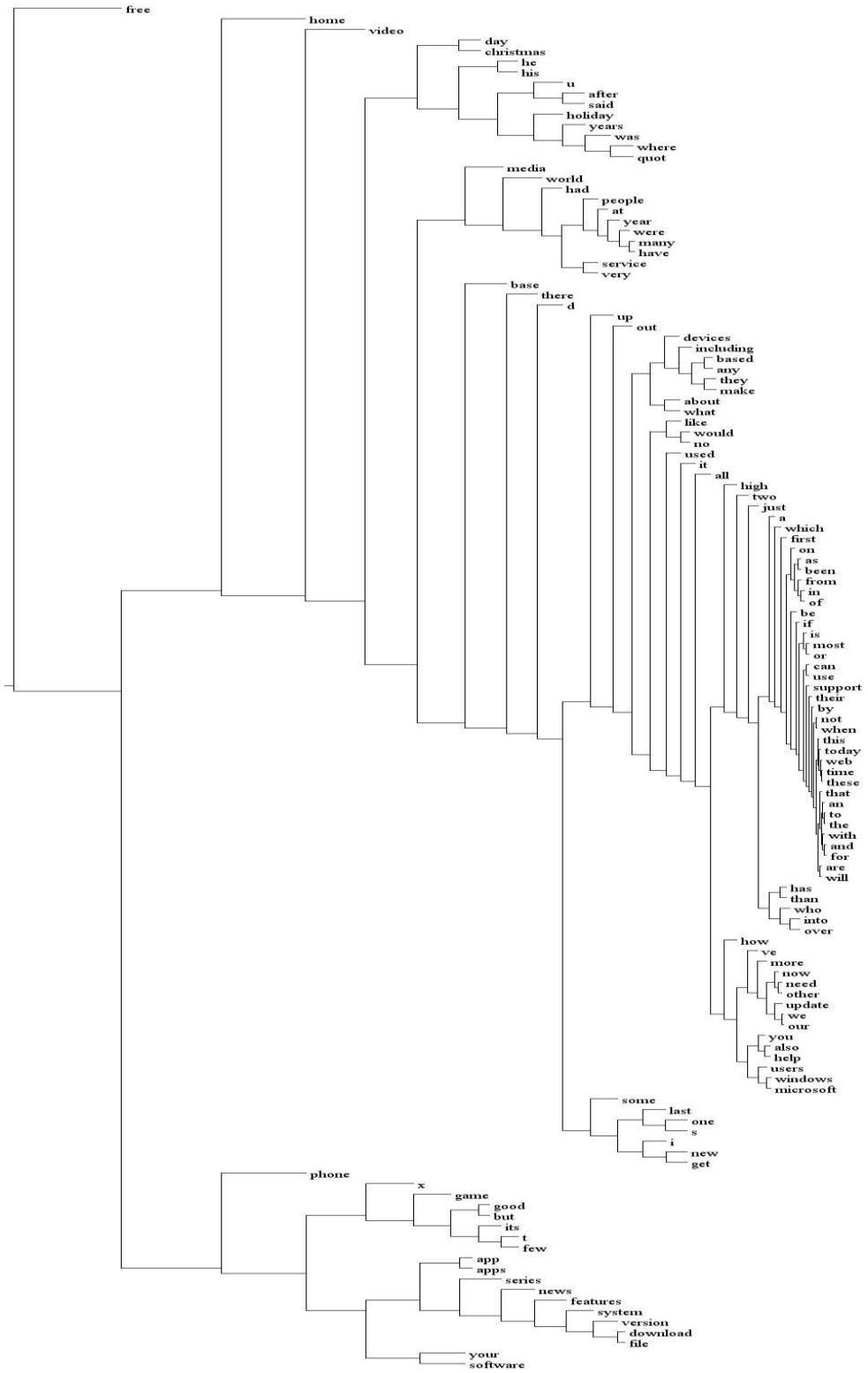

Figure 5. The Whole Dendrogram of all Words of 25 RSS

5. Conclusion

In this paper we illustrate how to cluster RSS data using hierarchical clustering method. Through clustering process analysis we describe the hierarchical clustering algorithm. In the experiment 25 RSS feeds are clustered. In order to experience the distances among different RSS feeds we also give the 2-D format and the relations of feeds are more clear in this format than in dendrogram. We cluster RSS contents by computing the distances of words with hierarchical clustering algorithm. Without modification of algorithm it can be used to cluster words by transposing matrix M. Usually the clustering of contents is have more meanings than clustering of words but the latter gives anther view and understanding of data.

Acknowledgements

This research is supported by Natural Science Foundation of Heilongjiang Province (F201034); Scientific Research Foundation for Doctor of Harbin University of Commerce (12DL024).

References

[1] http://en.wikipedia.org/wiki/RSS.
[2] P. Shi, "An Efficient Approach for Clustering Web Access Patterns from Web Logs", International Journal of Advanced Science and Technology, vol. 5, **(2009)**, pp. 1-13.
[3] M. Parimala, D. Lopez and N. C. Senthilkumar, "A Survey on Density Based Clustering Algorithms for Mining Large Spatial Databases", International Journal of Advanced Science and Technology, vol. 31, **(2011)**, pp. 59-66.
[4] W. Lei, W. Tongsen and Y. Ronghua, "Data Compression Algorithm based on Hierarchical Cluster Model for Sensor Networks", International Journal of Advanced Science and Technology, vol. 2, **(2009)**, pp. 72-84.
[5] T. Segaran and M. T. O'Brien, "Programming Collective Intelligence, O'Reilly Media, Sebastopol, **(2007)**.
[6] http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient.
[7] http://www.feedparser.org.

Authors

Wen Hu, master instructor, president of School of the Computer and Information Engineering, Harbin University of Commerce, backup leader of "Electronic Commerce" provincial key discipline echelon and academic leader of secondary doctoral discipline "electronic commerce and information service" in first-class doctoral discipline "business administration". His main research fields include computer network and communication, electronic commerce, embedded technology.

Qing he Pan, Doctor, Lecturer, teacher of School of the Computer and Information Engineering, Harbin University of Commerce. His main research fields include data analysis, electronic commerce, embedded technology.